D246–D251 *Nucleic Acids Research, 2013, Vol. 41, Database issue* *Published online 5 October 2012*
doi:10.1093/nar/gks915

LNCipedia: a database for annotated human lncRNA transcript sequences and structures

Pieter-Jan Volders[1], Kenny Helsens[2,3], Xiaowei Wang[4], Björn Menten[1], Lennart Martens[2,3], Kris Gevaert[2,3], Jo Vandesompele[1,*] and Pieter Mestdagh[1,*]

[1]Center for Medical Genetics, Ghent University, [2]Department of Medical Protein Research, VIB, [3]Department of Biochemistry, Ghent University, 9000 Ghent, Belgium and [4]Department of Radiation Oncology, Washington University School of Medicine, St. Louis, MO 63108, USA

Received August 15, 2012; Accepted September 10, 2012

ABSTRACT

Here, we present LNCipedia (http://www.lncipedia.org), a novel database for human long non-coding RNA (lncRNA) transcripts and genes. LncRNAs constitute a large and diverse class of non-coding RNA genes. Although several lncRNAs have been functionally annotated, the majority remains to be characterized. Different high-throughput methods to identify new lncRNAs (including RNA sequencing and annotation of chromatin-state maps) have been applied in various studies resulting in multiple unrelated lncRNA data sets. LNCipedia offers 21 488 annotated human lncRNA transcripts obtained from different sources. In addition to basic transcript information and gene structure, several statistics are determined for each entry in the database, such as secondary structure information, protein coding potential and microRNA binding sites. Our analyses suggest that, much like microRNAs, many lncRNAs have a significant secondary structure, in-line with their presumed association with proteins or protein complexes. Available literature on specific lncRNAs is linked, and users or authors can submit articles through a web interface. Protein coding potential is assessed by two different prediction algorithms: Coding Potential Calculator and HMMER. In addition, a novel strategy has been integrated for detecting potentially coding lncRNAs by automatically re-analysing the large body of publicly available mass spectrometry data in the PRIDE database. LNCipedia is publicly available and allows users to query and download lncRNA sequences and structures based on different search criteria. The database may serve as a resource to initiate small- and large-scale lncRNA studies. As an example, the LNCipedia content was used to develop a custom microarray for expression profiling of all available lncRNAs.

INTRODUCTION

Long non-coding RNAs (lncRNAs) constitute a recently discovered class of non-coding RNAs that grew in size drastically during the past few years. LncRNA genes give rise to long (>200 bp) and often multiexonic transcripts that are supposed not to get translated to protein, as commonly assessed by means of *in silico* prediction algorithms (1). In comparison with their protein-coding counterparts, lncRNA genes are poorly conserved (2) and are more numerous in biologically complex species (3). Although only a fraction of the lncRNA genes has been characterized experimentally, lncRNAs seem to function as transcriptional regulators through direct interaction with chromatin-modifying proteins and transcription factors (1,4,5).

LncRNAs with experimentally validated functions or expression patterns have been named accordingly. Notable examples are XIST (X inactive-specific transcript) (6), HOTAIR (HOX transcript antisense RNA) (7) and HULC (highly up-regulated in liver cancer) (8). The HUGO Gene Nomenclature Committee currently uses several schemes to name lncRNAs with an unknown function. LncRNAs that reside on the opposite strand to (antisense) or in an intron of (intronic) a protein-coding gene are named after the protein-coding gene with suffixes '-AS' and '-IT', respectively. Intergenic lncRNAs are numbered and get the prefix 'LINC' (9).

Recent advances in non-coding RNA research have led to the creation of several lncRNA resources. LncRNAdb focuses on lncRNA transcripts with well-described functions in literature (10), whereas the ncRNA database

*To whom correspondence should be addressed. Tel: +32 9 3326979; Fax: +32 9 3326549; Email: pieter.mestdagh@ugent.be
Correspondence may also be addressed to Jo Vandesompele. Tel: +32 479 353563; Fax: +32 9 3326549; Email: joke.vandesompele@ugent.be

(ncRNAdb) provides RNA sequences and annotation from different sources (11). The NONCODE database (12) contains a larger collection of human long non-coding RNAs (33 829) obtained from different sources and by different experimental procedures (13). Rfam provides structures and annotation of well-known RNA families along with predictions of new members of these families (14). However, it does not provide information for an individual lncRNA. Although each of these resources provides valuable information, database unification and integration of lncRNA transcript sequence details with a broad set of bioinformatics tools and a universal lncRNA gene building and naming scheme is currently lacking. Here, we present LNCipedia, a catalogue of 21 488 lncRNA transcripts that were clustered into genes and named accordingly, and they were analysed using multiple bioinformatics tools, revealing insights in lncRNA structure, experimentally verified (lack of) protein coding potential, function and regulation. We believe such a database facilitates human lncRNA research and communication among scientists.

DATABASE DEVELOPMENT

The sources used in the data collection step are listed in Table 1. The most recent version of each source at the time of development has been included. The sequences and annotations are extracted and stored in a mongoDB database using custom Perl scripts. To this purpose, import scripts for different file formats, such as FASTA, BED and GFF, have been developed. Redundant transcripts are grouped in a single record, while maintaining all annotation from the original sources. The web interface for LNCipedia is build using the Mojolicious Perl web framework and offers different ways of querying the data (Figure 1). LNCipedia will be updated when newer versions of the lncRNA sources are released or if new sources become available. In addition, researchers are encouraged to submit new transcript sequences or annotations trough lncipedia.org.

Of note, each of the input sources uses a different naming scheme. LncRNA researchers have previously used the gene symbol of the nearest protein coding gene to refer to a given lncRNA (15). Based on this strategy, we have implemented a universal lncRNA nomenclature to ease communication among researchers. Different lncRNA transcripts are considered to belong to the same gene if they share at least one (partially) overlapping exon and reside on the same DNA strand. In this way, transcripts are clustered into genes. These lncRNA genes are then named after the HUGO symbol of the nearest protein-coding gene on the same strand using the following scheme: 'lnc-HUGO-#'. The lncRNA genes are numbered, starting with the lncRNA gene closest to the protein-coding gene. A second number is added to denote the different transcript variants starting with the most upstream transcript, for example, lnc-MYCN-1:1 denotes transcript 1 from gene lnc-MYCN-1 (Figure 2).

Table 1. The different sources of lncRNA transcripts used for LNCipedia at the time of development[a]

Source	Version	Number of transcripts
Ensembl (biotype = lincRNA)	Version 64	9069
Human bodymap lincRNAs (2)		14 279
LncRNAdb (10)	September 2011	134
Total number of unique transcripts		21 488

[a]The database will be updated with new transcripts when new versions of the sources are released.

INTEGRATED ANALYSIS TOOLS

LncRNA-protein interactions are, in part, mediated by the secondary structure of the lncRNA. The Vienna RNA package (16,17) consists of a set of algorithms for predicting and analysing RNA secondary structures. We applied the RNAfold algorithm to generate a secondary structure plot and dot plot with pair probabilities. Both of these images are processed with the provided relplot.pl script to obtain a structure plot with colour annotated base pair probabilities. The output postscript (.ps) images are converted to the graphics interchange format (.gif) for display in web browsers.

Structural RNAs, such as miRNAs, have a significantly lower minimum free energy of folding compared with randomly shuffled sequences (18). The Randfold algorithm implements the randomization test and returns the mean free energy of folding and *P*-value for every RNA sequence. Hence, a significant *P*-value denotes a high propensity in the sequence towards a stable secondary structure.

Recently, it has been shown that lncRNAs can act as a miRNA sponge by binding specific microRNAs and, thus, interfering with their role as negative regulators of gene expression (5,19,20). We include miRNA seed predictions for every lncRNA to allow researchers to evaluate possible miRNA–lncRNA interactions. miRNA seed predictions were performed using the MirTarget2 algoritm (21).

PROTEIN CODING POTENTIAL

Assessment of protein coding potential is an important aspect in the study of non-coding RNAs. LNCipedia reports the outcome of two different protein coding potential prediction algorithms. The Coding Potential Calculator (CPC) applies a support vector machine classifier to the output of open reading frame analysis and Basic Local Alignment Search Tool search (22). CPC returns the predicted status of the transcript (coding/non-coding) and a coding potential score. We applied version 0.9 of the CPC software and report the predicted status and the coding potential score for every transcript. Another popular strategy for detection of

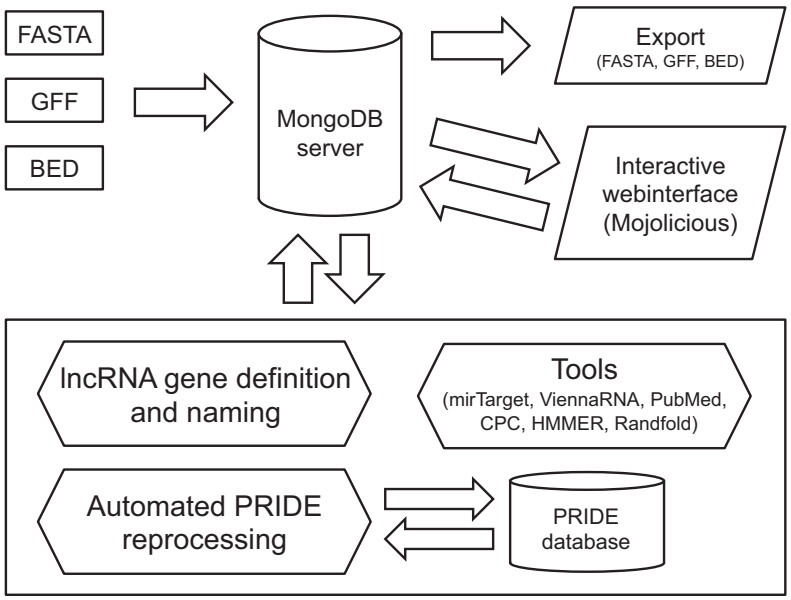


Figure 1. LNCipedia is generated in a multistep process that comprises importing, naming, analysis and visualization of lncRNA genes. Import scripts for the FASTA, BED and GFF file formats process lncRNA transcripts and detect redundancy. LncRNA naming is preceded by the creation of lncRNA transcript clusters and requires information on the nearest protein-coding gene on the same DNA strand. Every lncRNA transcript is subsequently analysed using multiple algorithms, and the results are appended to the database. A web-interface build using Perl enables lncRNA visualization and database querying.

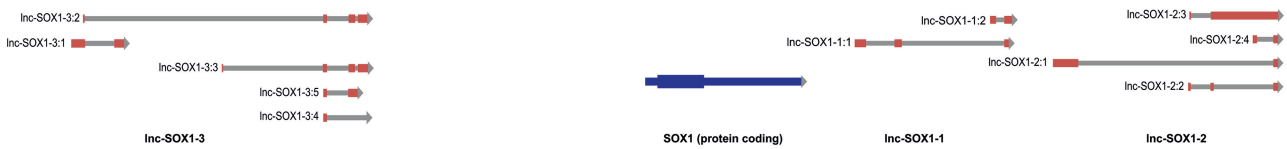


Figure 2. The SOX1 protein-coding gene locus contains three lncRNAs on the same DNA strand, numbered according to their distance in relation to SOX1. LncRNA transcripts are numbered according to their order in the gene, starting with the most upstream transcript.

coding sequences is based on known protein domains. The HMMER3 suite provides software based on hidden Markov models for sequence based homology searches (23). It is often used in combination with the Pfam protein families database (24). Using the hmmscan algorithm, we searched for Pfam protein domains in the RNA sequence. All six reading frames were translated *in silico*, and the number of hits in 5′ to 3′ and 3′ to 5′ direction are reported.

A unique feature of LNCipedia is the incorporation of an automated reprocessing pipeline that relies on publicly available fragmentation spectra from the PRIDE database at EMBL-EBI (25) to detect potentially coding lncRNAs. The concept behind this feature is that mass spectrometry based proteomics data may contain serendipitously recorded mass spectra derived from translated lncRNAs. As standard identification strategies in proteomics are based on searching these spectra against protein sequence databases, such as UniProtKB/Swiss-Prot (26), they are implicitly unable to detect coding forms of lncRNAs, as they are not present in these databases. To uncover such potential traces of coding lncRNAs, the spectra, thus, need to be re-searched against a purpose-built database that comprises a combination of the possible translations of known lncRNAs, the known proteins for that organism as obtained from a traditional sequence database and corresponding decoy sequences for both these constituent databases for quality control and FDR estimation purposes (27). A spectrum can, thus, be matched against a lncRNA, a known protein, or a decoy sequence. The known proteins must be included to prevent relatively low-scoring matches of spectra against lncRNAs to be picked up where a much better match for that spectrum can be found for a known protein.

We have implemented such a pipeline by using the SearchGUI tool (28) to run the X!Tandem (29) search algorithm. All results are then collated and filtered at 1% FDR by the PeptideShaker algorithm (http://code.google.com/p/peptide-shaker). The pipeline infers the original search parameters, such as mass errors and post-translational modifications both directly from the PRIDE database and by using the PRIDE automatic spectrum annotation pipeline (http://code.google.com/p/pride-asa-pipeline). All the tools and algorithms used are freely available as open source.

The pipeline has so far been ran on 149 PRIDE experiments from at least 15 different tissues, yielding 81 579 peptide-to-spectrum matches (PSMs) against the

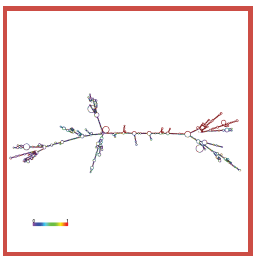
Incipedia.org

A comprehensive compendium of long non-coding RNAs | Home | Database | Search | Download | About | Contact

Transcript: lnc-SMUG1-3:6

Basic information

Incipedia transcript ID: lnc-SMUG1-3:6
Incipedia gene ID: lnc-SMUG1-3
Location: chr12:54356092-54368740
Strand: -
Transcript size: 2421 bp
Exons: 7
Sources: Ensembl release 64 - Sep 2011
Alternative transcript names: ENST00000424518
Alternative gene names: ENSG00000228630; HOTAIR



RNA sequence:

CCAGTTCTCAGGCGAGAGCCGCGGCTGACAGGGTCTGGGACAGAAGGAAAGCCCTCCAGCCTCCAGGCCCTGCCTTCTGCCTGCACATTCTGCCCTGATTTCCGGAACCTGGAAGCC
TAGGCAGGCAGTGGGGAACTCTGACTCGCCCTGTGCTCTGGAGCTTGATCCGAAAGCTTCCACAGTGAGGACTGCTCCGTGGGGGTAAGAGAGCACCAGGCACTGAGGCCTGGGAGTT
CCACAGACCAACACCCCTGCTCCTGGCGGCTCCCACCCGGGACTTAGACCCTCAGGTCCCTAATATCCCGGAGGTGCTCTCAATCAGAAAGGTCCTGCTCCGCTTCGCAGTGGAATG
GAACGGATTTAGAAGCCTGCAGTAGGGGAGTGGGGAGTGGAGAGAGGGAGCCCAGAGTTACAGACGGCGGCGAGAGGAAGGAGGGGCGTCTTTATTTTTTTAAGGCCCCAAAGAGT
CTGATGTTTACAAGACCAGAAATGCCACGGCCGCGTCCTGGCAGAGAAAAGCTGAAATGGAGGACCGGCGCCTTCCTTATAAGTATGCACATTGGCGAGAGAAGTGCTGCAACCTA
AACCAGCAATTACACCCAAGCTCGTTGGGGCCTAAGCCAGTACCGACCTGGTAGAAAAAGCAACCACGAAGCTAGAGAGAGAGCCAGAGGAGGGAAGAGAGCCGCCAGACGAAGGTG
AAAGCGAACCACGCAGAGAAATGCAGGCAAGGGAGCAAGCGCGGCAGTTCCCGGAACAAACGTGGCAGAGGGCAAGACGGGCACTCACAGACAGAGGTTTATGTATTTTTATTTTTTA
AAATCTGATTTGGTGTTCCATGAGGAAAAGGGAAAATCTAGGGAACGGGAGTACAGAGAGAATAATCCGGGTCCTAGCTCGCCACATGAACGCCCAGAGAACGCTGGAAAAACCTGA
GCGGGTGCCGGGGCAGCACCCGGCTCGGGTCAGCCACTGCCCCACACCGGGCCCACCAAGCCCCGCCCCTCGCGGCCACCGGGGCTTCCTTGCTCTTCTTATCATCTCCATCTTTAT
GATGAGGCTTGTTAACAAGACCAGAGAGCTGGCCAAGCACCTCTATCTCAGCCGCCGCCCGCTCAGCCGAGCAGCGGTCGGTGGGGGGACTGGGAGGCGCTAATTAATTGATTCCTTT
CGACTGTAAAATATCGCGGCCTCTACACGGAACCCATGACTCATAAACAATATATCTGTTGGGCGTGAGTGCACTGTCTCTCAAATAATTTTTCCATAGGCAAATGTCAGAGGGTTC
TGGATTTTTAGTTGCTAAGGAAAGATCCAAATGGGACCAATTTTAGCAGGCCCAAACAGAGTCCGTTCAGTGTCAGAAAATGCTTCCCCAAAGGGGTTGGGAGTGTCTTTTGTTGGAA
AAAAGCTTGGGTTATAGGAAAGCCTTTCCCTGCTACTTGTGTAGACCCAGCCCAATTTAAGAATTACAAGGAAGCGAAGGGTTGTGTAGGCCGGAAGCCTCTCTGTCCCGGCTGGAT
GCAGGGGACTTGAGCTGCTCCGGAATTTGAGAGGAACATAGAAGCAAAGGTCCAGCCTTTGCTTCGTGCTGATTCCTAGACTTAAGATTCAAAAACAAATTTTTAAAAGTGAAACCAG
CCCTAGCCTTTGGAAGCTCTTGAAGGTTCAGCACCCACCCAGGAATCCACCTGCCTGTTACACGCCTCTCCAAGACACAGTGGCACCGCTTTTCTAACTGGCAGCACAGAGCAACTCT

Structure:

Protein coding potential

CPC coding potential score: -1.19011 (noncoding) ?
HMMER Pfam domains in 3' to 5' reading frames: 0 ?
HMMER Pfam domains in 5' to 3' reading frames: 0

PRIDE database search

Number of hits in the PRIDE database: 0 ?

Secondary structure information

RNAfold image: download
Randfold minimum free energy: -825.83
Randfold P-value: 0.001

Targetting miRNAs

MirTarget2 predictions:

MicroRNA	MirTarget2 score ?
hsa-miR-3688-3p	93.51
hsa-miR-1251	87.25
hsa-miR-202-5p	82.56
hsa-miR-26b-3p	81.72
hsa-miR-892a	80.28

Available literature

- Guil et al., 2012
- Niinuma et al., 2012
- Kogo et al., 2011
- Schorderet et al., 2011
- Geng et al., 2011
- Kaneko et al., 2010
- Tsai et al., 2010
- Gupta et al., 2010

Figure 3. The transcript page in the web interface provides a clear overview of information available on a specific lncRNA transcript.

custom-built protein sequence database that includes UniprotKB/Swiss-Prot and LNCipedia translations (Supplementary Figure S1). Within these PSMs, there were just 14 matches that could provide evidence for translation of LNCipedia entries. However, after close inspection of the FDR of the PSMs that passed our quality criteria, we noticed that although the PSMs from UniProtKB/Swiss-Prot have an expected FDR of 0.9%, the subset of PSMs from translated LNCipedia entries comes with an overwhelming FDR of 166% (Supplementary Figure S2). As such, there are only vague suggestions so far that any of these entries can effectively be translated.

As the PRIDE database is growing exponentially, and additional lncRNA transcript discovery is ongoing, searches for potentially coding lncRNAs need to be carried out anew at regular intervals to stay up-to-date with the growing amount of public data. We, therefore, envision running the full pipeline on all applicable PRIDE data at a set interval of 3 months; thus, periodically updating the knowledge on which lncRNAs might have coding potential. The output of each reprocessing effort will be used to annotate the LNCipedia, and past results will be kept available as well.

Besides this recurrent re-analysis of the relevant publicly available proteomics data, we also plan to extend the statistical approach used to evaluate the identification of a lncRNA by including information about the consistency with which such an identification is found across (unrelated) PRIDE experiments. Indeed, a relatively poor match in any individual experimental data set that, however, keeps returning across many such data sets, may well be a real indication that translation is taken place for that lncRNA.

LNCIPEDIA ACCES

LNCipedia is publicly available through a web interface at http://www.lncipedia.org. The interface allows users to query lncRNAs by name, chromosomal region or (partial) sequence. Several statistics are calculated that allow the user to evaluate different parameters regarding lncRNA secondary structure and regulation (Figure 3). The entire LNCipedia collection is available for download in the FASTA, GFF or BED format.

LncRNA researchers can contribute to LNCipedia by contacting the authors. In addition, registered users can modify existing records (updating aliases and adding PubMed literature records) directly using a web interface.

LNCRNA EXPRESSION ARRAY

The LNCipedia content can prove useful when designing large-scale screening experiments, such as lncRNA gene expression profiling. As a proof of concept, we have developed a custom lncRNA gene expression array using the Agilent Sureprint 60 k platform. In addition to roughly 33 000 probes for protein coding genes, we selected 23 042 probes for lncRNA transcripts in LNCipedia covering 97% of all LNCipedia transcripts with at least one probe (Agilent MicroArray Design ID: 039714). The performance of the expression array was evaluated using RNA sample titrations according to the MicroArray Quality Control standards (30). Adequate titration response of the lncRNA probes is shown in Supplementary Figure S3.

CONCLUSION AND FUTURE DIRECTION

Three important features are unique to LNCipedia: gene definitions and usage of a universal nomenclature for lncRNA transcripts, PRIDE analysis for detection of lncRNAs that may code for small peptides and miRNA seed predictions for lncRNA transcripts. These, along with the other tools available, are expected to make LNCipedia a powerful resource for human lncRNA research.

With the advances in RNA sequencing technology, more lncRNA genes are expected to get discovered. The authors will update LNCipedia when new sequences are reported in the literature or in other sources. In addition, new features will be developed to increase the interactive capabilities of LNCipedia. In this way, the lncRNA community will be able to upload and maintain records in the database. LNCipedia has the potential to become a community resource for lncRNA transcript information and annotation.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Figures 1–3 and Supplementary Methods.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the equal contribution of J.V. and P.M.

FUNDING

Ghent University Multidisciplinary Research Partnership ‘Bioinformatics: from nucleotides to networks’ (to P.J.V., L.M., K.G., J.V.); National Institutes of Health [R01GM089784 to X.W.]; Flemish Fund for Scientific Research Flanders (FWO) (to P.M.); Ghent University Special Research Fund (BOF) (to J.V.). Funding for open access charge: Ghent University.

Conflict of interest statement. None declared.

REFERENCES

1. Mercer,T. and Dinger,M. (2009) Long non-coding RNAs: insights into functions. *Nat. Rev. Genet.*, **10**, 155–159.
2. Cabili,M.N., Trapnell,C., Goff,L., Koziol,M., Tazon-Vega,B., Regev,A. and Rinn,J.L. (2011) Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev.*, **25**, 1915–1927.
3. Taft,R.J. and Mattick,J.S. (2003) Increasing biological complexity is positively correlated with the relative genome-wide expansion of non-protein-coding DNA sequences. *Genome Biol.*, **5**, P1–P24.
4. Wang,K.C. and Chang,H.Y. (2011) Molecular mechanisms of long noncoding RNAs. *Mol. Cell*, **43**, 904–914.

5. Guttman,M. and Rinn,J.L. (2012) Modular regulatory principles of large non-coding RNAs. *Nature*, **482**, 339–346.
6. Brown,C.J., Ballabio,A., Rupert,J.L., Lafreniere,R.G., Grompe,M., Tonlorenzi,R. and Willard,H.F. (1991) A gene from the region of the human X inactivation centre is expressed exclusively from the inactive X chromosome. *Nature*, **349**, 38–44.
7. Gupta,R.A., Shah,N., Wang,K.C., Kim,J., Horlings,H.M., Wong,D.J., Tsai,M.C., Hung,T., Argani,P., Rinn,J.L. *et al.* (2010) Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. *Nature*, **464**, 1071–1076.
8. Panzitt,K., Tschernatsch,M.M., Guelly,C., Moustafa,T., Stradner,M., Strohmaier,H.M., Buck,C.R., Denk,H., Schroeder,R., Trauner,M. *et al.* (2007) Characterization of HULC, a novel gene with striking up-regulation in hepatocellular carcinoma, as noncoding RNA. *Gastroenterology*, **132**, 330–342.
9. Wright,M.W. and Bruford,E.A. (2011) Naming 'junk': human non-protein coding RNA (ncRNA) gene nomenclature. *Hum. Genomics*, **5**, 90–98.
10. Amaral,P.P., Clark,M.B., Gascoigne,D.K., Dinger,M.E. and Mattick,J.S. (2010) lncRNAdb: a reference database for long noncoding RNAs. *Nucleic Acids Res.*, **39**, D146–D151.
11. Szymanski,M., Erdmann,V.A. and Barciszewski,J. (2007) Noncoding RNAs database (ncRNAdb). *Nucleic Acids Res.*, **35**, D162–D164.
12. Liu,C., Bai,B., Skogerbø,G., Cai,L., Deng,W., Zhang,Y., Bu,D., Zhao,Y. and Chen,R. (2005) NONCODE: an integrated knowledge database of non-coding RNAs. *Nucleic Acids Res.*, **33**, D112–D115.
13. Bu,D., Yu,K., Sun,S., Xie,C., Skogerbo,G., Miao,R., Xiao,H., Liao,Q., Luo,H., Zhao,G. *et al.* (2011) NONCODE v3.0: integrative annotation of long noncoding RNAs. *Nucleic Acids Res.*, **40**, D210–D215.
14. Gardner,P.P., Daub,J., Tate,J., Moore,B.L., Osuch,I.H., Griffiths-Jones,S., Finn,R.D., Nawrocki,E.P., Kolbe,D.L., Eddy,S.R. *et al.* (2010) Rfam: wikipedia, clans and the 'decimal' release. *Nucleic Acids Res.*, **39**, D141–D145.
15. Guttman,M., Amit,I., Garber,M., French,C., Lin,M.F., Feldser,D., Huarte,M., Zuk,O., Carey,B.W., Cassady,J.P. *et al.* (2009) Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature*, **458**, 223–227.
16. Hofacker,I.L. (2003) Vienna RNA secondary structure server. *Nucleic Acids Res.*, **31**, 3429–3431.
17. Hofacker,I.L., Fontana,W., Stadler,P.F., Bonhoeffer,L.S., Tacker,M. and Schuster,P. (1994) Fast folding and comparison of RNA secondary structures. *Monatsh. Chem. Chem. Mon.*, **125**, 167–188.
18. Bonnet,E., Wuyts,J., Rouzé,P. and Van de Peer,Y. (2004) Evidence that microRNA precursors, unlike other non-coding RNAs, have lower folding free energies than random sequences. *Bioinformatics*, **20**, 2911–2917.
19. Cesana,M., Cacchiarelli,D., Legnini,I., Santini,T., Sthandier,O., Chinappi,M., Tramontano,A. and Bozzoni,I. (2011) A long noncoding RNA controls muscle differentiation by functioning as a competing endogenous RNA. *Gastroenterology*, **147**, 358–369.
20. Kretz,M., Webster,D.E., Flockhart,R.J., Lee,C.S., Zehnder,A., Lopez-Pajares,V., Qu,K., Zheng,G.X., Chow,J., Kim,G.E. *et al.* (2012) Suppression of progenitor differentiation requires the long noncoding RNA ANCR. *Genes Dev.*, **26**, 338–343.
21. Wang,X. and El Naqa,I.M. (2008) Prediction of both conserved and nonconserved microRNA targets in animals. *Bioinformatics*, **24**, 325–332.
22. Kong,L., Zhang,Y., Ye,Z.Q., Liu,X.Q., Zhao,S.Q., Wei,L. and Gao,G. (2007) CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic Acids Res.*, **35**, W345–W349.
23. Eddy,S.R. (2011) Accelerated profile HMM searches. *PLoS Comput. Biol.*, **7**, e1002195.
24. Punta,M., Coggill,P.C., Eberhardt,R.Y., Mistry,J., Tate,J., Boursnell,C., Pang,N., Forslund,K., Ceric,G., Clements,J. *et al.* (2011) The Pfam protein families database. *Nucleic Acids Res.*, **40**, D290–D301.
25. Martens,L., Hermjakob,H., Jones,P., Adamski,M., Taylor,C., States,D., Gevaert,K., Vandekerckhove,J. and Apweiler,R. (2005) PRIDE: the proteomics identifications database. *Proteomics*, **5**, 3537–3545.
26. Sadygov,R.G., Cociorva,D. and Yates,J.R. (2004) Large-scale database searching using tandem mass spectra: looking up the answer in the back of the book. *Nat. Methods*, **1**, 195–202.
27. Vaudel,M., Burkhart,J.M., Sickmann,A., Martens,L. and Zahedi,R.P. (2011) Peptide identification quality control. *Proteomics*, **11**, 2105–2114.
28. Vaudel,M., Barsnes,H., Berven,F.S., Sickmann,A. and Martens,L. (2011) SearchGUI: an open-source graphical user interface for simultaneous OMSSA and X!Tandem searches. *Proteomics*, **11**, 996–999.
29. Craig,R. and Beavis,R.C. (2004) TANDEM: matching proteins with tandem mass spectra. *Bioinformatics*, **20**, 1466–1467.
30. Canales,R.D., Luo,Y., Willey,J.C., Austermiller,B., Barbacioru,C.C., Boysen,C., Hunkapiller,K., Jensen,R.V., Knight,C.R., Lee,K.Y. *et al.* (2006) Evaluation of DNA microarray results with quantitative gene expression platforms. *Nat. Biotechnol.*, **24**, 1115–1122.